MTI ground data processing and science retrieval pipeline architecture

John J. Szymanski,[*] Christoph C. Borel, William B. Clodius, Anthony B. Davis,
Christopher E. Heil, James B. Krone, Keri Ramsey, Barham W. Smith, Kathy Smith,
James Theiler, Pierre V. Villeneuve and Paul G. Weber

Los Alamos National Laboratory
Mail Stop C323, Los Alamos, NM 87545

ABSTRACT

The major science goal for the Multispectral Thermal Imager (MTI) project is to measure surface properties such as vegetation health, temperatures, material composition and others for characterization of industrial facilities and environmental applications. To support this goal, this program has several coordinated components, including modeling, comprehensive ground-truth measurements, image acquisition planning, data processing and data interpretation. Algorithms have been developed to retrieve a multitude of physical quantities and these algorithms are integrated in a processing pipeline architecture that emphasizes automation, flexibility and robust operation. In addition, the MTI science team has produced detailed site, system and atmospheric models to aid in system design and data analysis. This paper will provide an introduction to the data processing and science algorithms for the MTI project. Detailed discussions of the retrieval techniques will follow in papers from the balance of this session.

Key words: Data processing centers, multispectral analysis, remote sensing algorithms, thermal infrared

1. INTRODUCTION AND GOALS

The overall goals for the MTI mission are discussed by Weber et al.[1] In summary, the MTI mission is to demonstrate the efficacy of highly accurate space-based multispectral imaging with state-of-the-art calibration and modest spatial resolution for passive characterization of industrial facilities and related environmental impacts. MTI provides simultaneous data for atmospheric characterization at good spatial resolution to aid in this analysis. Additionally, MTI can be used for local environmental monitoring and other civilian applications. Substantial effort was put into end-to-end modeling to guide the sensor design, the detailed space sensor engineering, and development of analysis techniques to form a balanced, self-consistent mission. Assembly of the MTI satellite nears completion, and is scheduled for launch in late 1999. The ambitious goals for MTI drove the design of the sophisticated payload and advanced calibration systems, which are the subject of the first set of MTI papers at this conference, as well as driving the design of the data processing and analysis tools that are described in the second set of papers.

The major components of the MTI experimental program include:
1. In-depth site, system and atmospheric modeling
2. A broad set of science retrieval algorithms
3. Detailed calibration and registration
4. The data processing center infrastructure
5. Ground-truth experimentation.

Several of these topics are discussed in detail in other papers contained in this volume. This paper provides an overall summary of the program and the data products, and details the data processing system and related facilities. Other talks in the session will detail (1) The science retrievals (Borel et al.[2]), (2) An experimental approach to MTI retrievals using genetic programming (Theiler et al.[3]), (3) Materials identification and anomaly detection with MTI (Stallard and Taylor[4]), and (4) Data and algorithm validation with ground-truth campaigns (Garrett et al.[5]). Previously, Clodius presented details about the

[*]Correspondence: Email: szymanski@lanl.gov; Telephone: 505 665 9371; Fax: 505 667 3815

MTI on-orbit calibration, which is also closely related to the experimental program (Clodius et al.[6]).

2. DPAC FACILITIES

The central repository for algorithms, data, metadata, and general analysis tools is the MTI Data Processing and Analysis Center (DPAC) located at Los Alamos National Laboratory (LANL). The DPAC processes all raw MTI data to produce several levels of data products. The MTI data product levels follow the convention presented in the *Earth Observing System* (EOS) *Reference Handbook*.[7] Alexander gives an alternative categorization of data product levels that was available after the MTI data products were defined.[8] What the EOS conventions mean in practice for MTI is described in Section 2.1. The processing pipeline that produces the data products is described in section 2.2.

A database is used to request automated processing and to monitor and store the results of this processing (section 2.3). The DPAC maintains hardware, described in section 2.4, to process and store MTI imagery and related data, as well as facilities to search for and retrieve this information. All of the hardware and much of the software are based on commercial off-the-shelf (COTS) and open-source software components and is largely platform-independent. Internal and external users have access to the database and thumbnail images through a web browser interface (see section 2.5).

MTI data is organized into image cubes (the term "image" is also used in this paper) and sequences of image cubes, where an image cube is a series of two-dimensional arrays containing the radiance information for a given source taken in different spectral bands. Because MTI is capable of taking multiple ground acquisitions of a single target area during a single overhead pass, and because MTI images are accompanied by several calibration data acquisitions, image cubes are organized into sequences. For example, a typical image sequence might consist of the components shown in Fig. 1. The image sequence flag indicates the beginning of a sequence, whic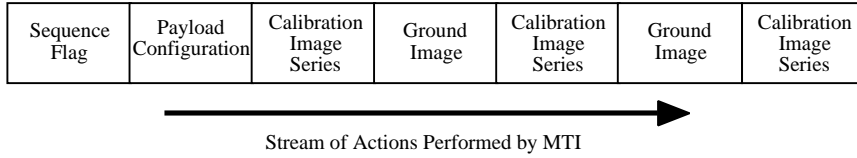h is followed by payload configuration information. Several calibration acquisitions are taken, followed by a ground image, more calibration images, a ground image and more calibration images. The number and type of calibration images is discussed in detail by Clodius et al.[6] Typical calibration images including high and low temperature black bodies, white body, and retro mirror (a mirror that reflects the radiance emitted from or reflected by the focal plane back onto itself) on-board calibration sources, as well as occasional calibration images taken of external sources.[9] The external calibration images will be acquired separately from ground image acquisitions.

Figure 1. An image sequence. The raw data file will have data packets that are in an arbitrary order. This figure is an illustration of the commanded sequence of events for an image sequence.

Table 1. MTI image types.

Image Type	Description
ada	Data acquired with the Aperture Door Assembly in place (a black-painted surface)
bb1	Data acquired with black body 1 flooding the focal plane
bb2	Data acquired with black body 2 flooding the focal plane
rem	Data acquired with retro mirror at the intermediate pupil
lp1	Data acquired with VNIR lamp 1 flooding the focal plane
lp2	Data acquired with VNIR lamp 2 flooding the focal plane
gnd	Data acquired with the door open and the satellite oriented toward the earth
spc	Data acquired with the door open and the satellite oriented away from the earth (spc stands for space)
vrp	Data acquired with the visible reflectance panel reflecting sunlight into the external aperture
moo	Data collected while the satellite is pointed at the moon

To easily sort image files and to identify images in the databases, image types have been defined as shown in Table 1. Deep space images are done for calibration (to measure the zero-radiance system offset) as well as possibly for astrophysical measurements. Moon images are used for calibration, as are all other image types except for ground images.

Some of the ground images will be used, in addition, for validation of algorithms and sensor performance (see section 5).

All MTI data products are stored in the Hierarchical Data Format, or HDF, with most ancillary data available in the same file as the image data.[10] This format allows the DPAC software to build data structures within the data file that match the data

Table 2. MTI data product levels. These categories were taken from the *EOS Reference Handbook*.

Data Level	Description
Level 0	instrument/payload data at full resolution: any and all communication artifacts are removed. For MTI, this data level consists of re-formatted and re-ordered images built up from individual lines from the push broom scan.
Level 1A	Decompressed, reconstructed, and unprocessed instrument/payload data at full resolution, time referenced and annotated with ancillary information including geometric and radiometric calibration coefficients, and georeferencing parameters computed and appended, but not applied, to Level 0 data. The DPAC will produce the Level1A-base data product with these qualities.
Level 1B	Level 1A data that has been processed to sensor units. MTI has several products in this category corresponding to registered or unregistered data as well as two levels of geolocation.
Level 2	Derived geophysical quantities. MTI will produce a number of these products that are described in the paper by Borel et al. (1999).
Level 3	Geophysical quantities mapped on uniform space-time grid scales. No specific products have yet been defined in this category for MTI, but the experimental program will have several sites to be imaged periodically – Level 3 data products will likely be defined for some geophysical quantities measured for these sites.
Level 4	Model output or results from analyses of lower-level data. MTI will have several of these (e.g., vegetation health and subpixel temperature analyses).

Table 3. The MTI Level 0 and 1 data products.

Product ID	Data Set Name	Notes
Level 0-Raw	Archival Data	
Level 0-DN	DN images in Logical Image Structure (LIS) format (all bands)	Subset of Level 0-Raw
Level 0-SOH	State of Health database	Subset of Level 0-Raw
Level 1A-Base	Annotated look data	Standard. Adds calibration and pointing correction data.
Level 1B-U	Unregistered but calibrated Top of Atmosphere (TOA) radiances in LIS format (all bands)	Standard. Level 1A-Base with calibration applied.
Level 1B-Att	Spacecraft orbit and attitude solutions	Standard
Level 1B-R-QL	Quick Look TOA data for browsing	Standard. Degraded resolution registered data.
Level 1B-R-Coreg	Co-registered TOA radiance cube at sensor	Standard; bands resampled to common grid.
Level 1B-R- Geo	Co-registered Geo-located TOA radiance cube	Standard; resampled to geo-referenced coordinates.
Level 1B-R- Topo	Topographically co-registered and geo-located TOA radiance cube	Non-standard

structures used in the pipeline. Presently the HDF-EOS extensions are not used in the DPAC. It is expected that the National Imagery Transmission Format (NITF) will also be supported by the DPAC. Other formats may be supported as special requests (i.e., supported as resources permit).

2.1 MTI data products and data product levels

The MTI data products are categorized according to the *EOS Reference Handbook* as shown in Table 2.[7] Level 0 and 1 products are discussed further in this paper and Level 2 products are the subjects of the paper by Borel et al.[2] The complete list of level 0 and1 data products is in Table 3. Level 2 products are listed in Table 4 and Level 4 products are listed in Table 5.

Table 4. MTI level 2 data products.

Product ID	Data Set Name	Notes
Level 2-WM	Water Mask	Standard
Level 2-CirM	Cirrus Mask	Standard daytime
Level 2-CldM	Non-Cirrus Cloud Mask	Standard daytime
Level 2-Vap	Atmospheric Water Vapor Image	Standard daytime.
Level 2-WST-$\alpha^\dagger$	Water Surface Temperature	Standard. The optional "-α" suffix indicates the retrieval algorithm.[†]
Level 2-LST	Land Surface Temperature	Non-standard.
Level 2-Refl	Surface Reflectance cube corrected for atmosphere	Non-standard daytime.
Level 2-NDVI	The Normalized Difference Vegetation Index	Standard daytime
Level 2-Mat	Material Identification map	Non-standard daytime
Level 2-PC	Principal Components cube	Standard
Level 2-MNF	Minimum Noise Fraction cube	Standard
Level 2-IW	Iterative Wiener Filtered radiance cube	Non-standard
Level 2-MEM	Maximum Entropy Method filtered radiance cube.	Non-standard
Level 2-Pixon	Pixon based radiance cube	Non-standard

[†]α has the form M-L-T, with the allowed substitutions for M (method) being either R (robust) or P (physics), for L (looks) the numerals 1 or 2, and for T (time) either day or night.

Table 5. MTI level 4 data products.

Product ID	Data Set Name	Notes
Level 4-ALGE	ALGE comparison data for water cooling systems	Non-standard
Level 4-RAMS	RAMS comparison data for cooling tower	Non-standard
Level 4-Subpix	Subpixel temperature analysis	Non-standard
Level 4-Veg	Vegetation Health Status map	Non-standard

2.2 The DPAC automated processing pipeline

The DPAC automated pipeline is designed to produce all Level 0 products and most of Level 1. One exception is geolocation, which is done at Level 1 and is not presently automated. Parts of Level 2 and any future Level 3 products will also be automated. The Level 2 science retrievals need to be flexible in both their order of execution and for their input parameters. The automated pipeline structure is quite flexible and can handle most of these needs, although the different modes of running Level 2 will not be completely determined until after launch, when more experience is gained with the MTI instrument and data. The architecture supports the definition and automatic updating of Level 3 products when they are

defined.

A high-level block diagram of the automated pipeline is shown in Fig. 2. The major features include: (1) Automatic processing of data through Level 0 by simply placing a raw data file in the appropriate location. No user intervention is necessary for this processing to occur. (2) Automatic generation of a Level 1 (and higher) request when Level 0 is complete. (3) Input from the database to guide the processing, as well as output to the database to monitor progress and post results. (4) All the facilities needed to run the data multiple times (e.g., when new versions of processing algorithms are developed). A "rid" is a request id, which is used to track data processed on multiple occasions.

2.3 Databases

The DPAC maintains two databases, one for DPAC (as opposed to spacecraft) operations information and another to store spacecraft and payload state-of-health data. The operations database provides input to the processing (automated or otherwise), monitors progress of the processing, and stores results of the processing (including some summary Level 2 outputs, such as the fractional cloud cover of an image). This database provides flexibility and extensibility to the automated processing pipeline. The operations database tables are shown in Table 6. The state-of-health database provides a convenient means for accessing state-of-health data through general queries for trend analyses as well as providing direct input to Level 1 and 2 retrievals. Both databases use the Informix Standard Edition™ database engine running under Linux.
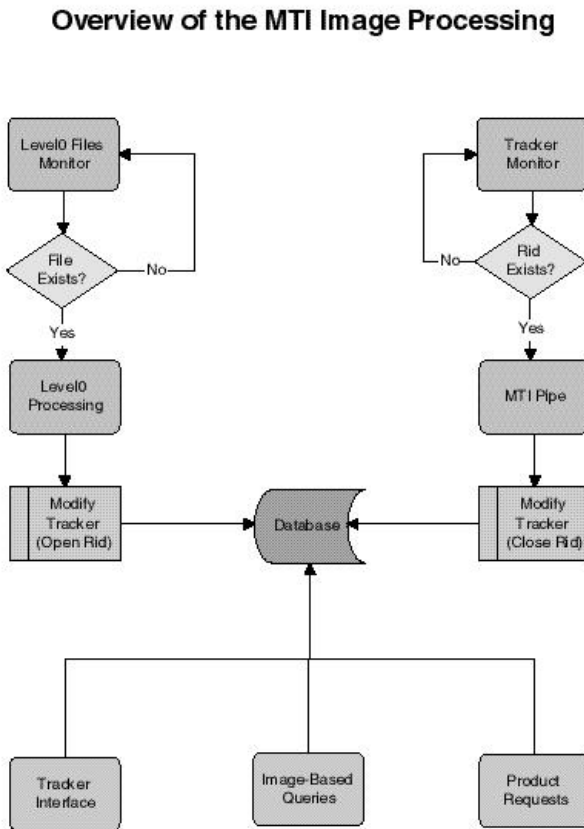
The image metadata stored in the operations database is used to identify images of interest to researchers both within and outside Los Alamos National Laboratory. This portion of the database has information relevant to the operation of the payload (e.g., what portions of the focal plane are active, integration times, the total number of scan lines, etc.), to the location imaged (e.g., latitude and longitude of the image, look angle, etc.), and to the description of the image (e.g., the fractional cloud cover, the fraction of the scene identified as water, average atmospheric water vapor, etc.). A web browser interface gives access to the metadata, allowing researchers to query on many combinations of parameters to select images of interest (see section 2.5).

Figure 2. The automated pipeline processing for the MTI DPAC. The term rid stands for a request id. Note the central role the database plays in automating the processing

2.4 Hardware and software standards and architecture

The DPAC hardware is based on Pentium™ PCs running the Linux operating system. The hardware was chosen because of its good price/performance in comparison to standard graphics workstations. The DPAC runs a total of approximately 20 PCs consisting of a mixture of Pentium II PCs running at 266 to 450 MHz and Pentium III PCs running at 500 MHz. Most machines are dual-processor, with 0.5 to 1 GB of memory, at least 9 GB or disk storage, and provide quite good performance at highly competitive prices. Six monitorless workstations are run in a compute-server cluster networked together with 100 Mbit ethernet. All raw and processed data products are archived on a mass storage system capable of storing hundreds of terabytes. The mass storage system is composed of 250 GB of local disk cache providing fast access to recently used data.

Less recently used data is stored near-line in automated tape libraries. From the user's point of view, all files appear to be stored locally. The Linux operating system is an open-source operating system that gives good performance, security, rapidly growing commercial support, and low cost. Processing occurs in parallel on several machines at a time. No attempt was made to parallelize individual codes with Message Passing Interface or Parallel Virtual Machine – the (course-grained) parallelization occurs by running different portions of the pipeline on different machines, as well as by running different data sets on different machines.

Table 6. Tables in the MTI operations database. The image tables are used to store information relevant to individual images or raw data files. The processing tables store information used to control DPAC operations.

Operations Database Tables	
Image tables	**Description**
External_looks	Detailed information about every external image.
Images	Basic information about every image processed (including calibration and external images).
Image_files	Raw data file information.
Image_types	Lookup table for image types.
Level0_files	Detailed information about every raw file processed.
Seq_summary	Basic information about every sequence processed based on Sequence ID
Processing tables	**Description**
Control_point	Information about control points used in the generation of the Level2_Geo product.
Geoloc_detail	Information required for the generation of the Level2_Geo product.
Pipe_paths	Information about the location of output products based on date and software version.
Product_prereq	Prerequisite products for all products.
Products	Products to be generated as part of the MTI pipeline.
Tracker	Requests for products to be generated.
Tracker_files	Non-standard output for product requests.
Tracker_prods	Products requested as part of the request specified in 'tracker'.

The DPAC system is based on standard COTS and open-source software components, including IDL™, PERL, Java, bash shell scripts and some C code. The majority of the software is written in IDL, because of its inherent strengths in image processing and rapid prototyping capability. Object-oriented IDL was used for input/output for ease of maintenance. PERL, Java or C code are used in places where performance or software design considerations dictate. In addition, the DPAC uses some standard simulation packages from other institutions that are written in FORTRAN, such as MODTRAN.

2.5 Web browser access to images and metadata

A web browser is used to access the operations database and search for images of interest. The user can select images in one of three ways: (1) From a list of sites that are
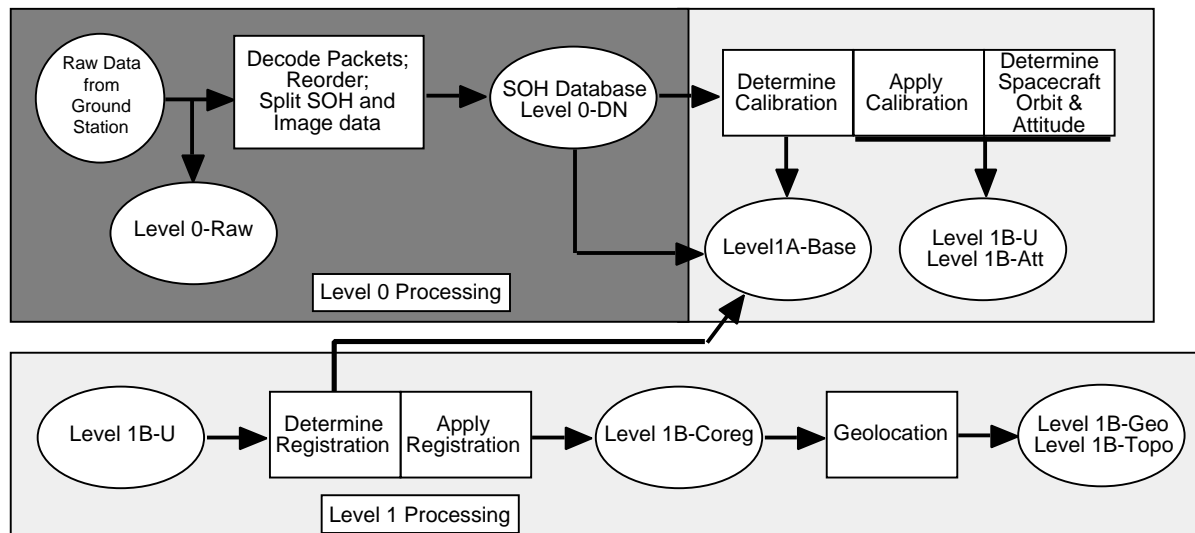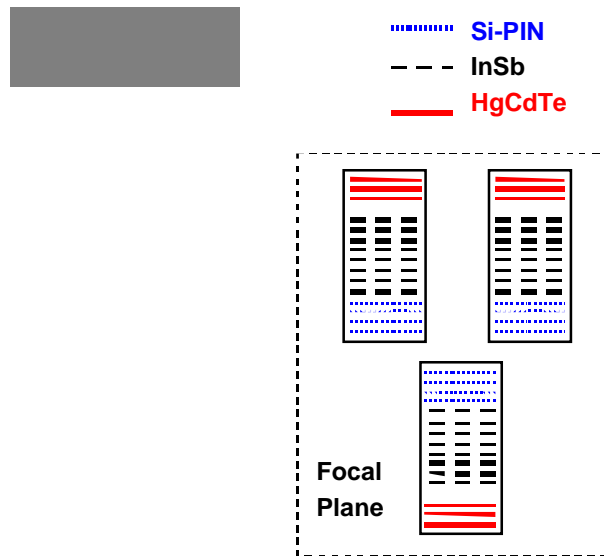

Figure 3. Level 0 and Level 1 processing steps. At all stages, the operations and state-of-health databases are accessed and/or updated as needed.

periodically imaged by MTI. (2) By clicking on maps that show sites imaged by MTI. (3) By doing a general query on the operations database. An external user can use this interface to order data products of interest and to request data products that have not yet been produced for this ground image (this is considered a special request that the DPAC will handle resources permitting). The web interface is the means for principal investigators to inspect the imagery and track the processing of their images. This software package is written in Java, using Remote Method Interface and Java Database Connectivity Package to interface with the database.

3. LEVEL 0 AND 1 PROCESSING

The Level 0 and Level 1 processing is shown in Fig. 3. All of the Level 0 and most of the Level 1 data products are produced automatically. The automated processing starts when a raw data file is transferred from the ground station, which happens on average twice each day at ~2:00 a.m. and ~2:00 p.m. (approximately one hour after the downlink).

Level 0 products correspond to a re-ordering and re-formatting of the raw data. For MTI, the image data are taken as scan lines within a single spectral band. To cover a given line on the ground at full swath width, three linear detector arrays need to be read out (see Fig. 4), although the arrays do not pass over the same section of ground at the same time. The Level 0 processing places scan lines from the same Sensor-Chip Assembly (SCA) and same band together in a data structure within an HDF file, which forms an uncalibrated and unregistered, but nevertheless useful, image. All image data for all the bands active for a given image, as well as all active SCAs are placed in a single file. One or more external images (typically taken of the earth's surface, although astrophysical scenes may also be taken) are part of an image sequence that also includes several calibration images, taken with either on-board or external sources.

The state-of-health database is populated at the beginning of Level 0 processing. Further use of these data is made by the Level 1 and 2 retrievals, which need state-of-health quantities to determine the instrument response.

Level 1 products involve detailed calibration and registration of the data. The calibration procedures are covered by Clodius et al.[6] The registration has been presented elsewhere, but a short discussion is included here.[11] The basic problem is that data are acquired in different bands at different times for the same point on the ground. After calibration the data in the different bands is registered together and resampled. The need for resampling and registration is compounded by the fact that the spacecraft is not perfectly steady. The radiance correlation between bands is used to determine a "jitter function", which describes the spacecraft roll and pitch as a function of time. The registration also removes optical distortions and connects

together the sub-images produced by the three SCAs.

The final step in the Level 1 processing is to geolocate each external image. This procedure is done by hand using the ENVI (ENvironment for Visualizing Images) image-processing package (™Better Solutions Consulting, LLC). The databases support geolocation by providing for each image a list of control point descriptions, latitudes and longitudes. Control point locations will be provided by GPS for the ground truth experimentation sites. Standard geolocation will be done to a uniform height above the WGS84 ellipsoid. For selected scenes geolocation will be done using a digital elevation model.

4. LEVELS 2, 3 & 4 SCIENCE RETRIEVALS

The Level 2 retrievals are covered elsewhere in this volume.[2] Level 2 is where the bulk of geophysical quantities needed for remote sensing research are retrieved from the imagery. The pipeline will be set up to perform Level 2 retrievals in several different sequences, as well as with varying parameters. The reason is that different scenes will likely require differences in the Level 2 (and perhaps Level 1) processing. Indeed, much of the non-routine (non-automated) analysis of imagery may be done using a highly-flexible image browser.

The processing pipeline is setup to accommodate Level 3 products by automatically including Level 2 results in a time-series product. This is particularly useful for sites that are periodically imaged by MTI, and where trending data is relevant.

At the time of this writing three types of Level 4 products are anticipated: (1) Comparisons with facility models, (2) Determination of vegetation health, and (3) Subpixel temperature retrieval.[12] A Level 4 product is a model output or result from analyses of lower-level data, implying substantial user interaction. Thus, these products will not be produced for all external images.

5. GROUND-TRUTH CAMPAIGNS

The MTI ground-truth (or Validation and Verification, or V&V) program is lead by the Savannah River Technology Center and is covered in detail elsewhere in this volume.[5] This section contains a discussion of the relation of the ground-truth effort to the rest of the experimental program.

The MTI V&V program consists of several aspects: (1) Selection of appropriate sites to verify all the MTI measurement capabilities. (2) Development or procurement of instrumentation to directly measure geophysical quantities retrieved from MTI data. (3) Collaboration with organizations managing the verification sites to share data. (4) Facility modeling to interpret the results. The DPAC will archive all meteorological, facility-provided and direct measurements relevant to MTI V&V. In some cases, meteorological measurements are needed for many days before an MTI V&V acquisition, so complete time-history information will be archived.

Sites are selected to provide complete V&V of possible MTI applications. Thus, sites including features that directly measure instrument performance characteristics have been chosen, such as long, straight objects to evaluate inter-band registration performance and line-spread functions. Calibrated reflectance panels will also be used. In addition, sites have been selected to validate specific MTI capabilities, such as Crater Lake or cooling lakes to get cold and hot water temperatures, respectively. Sites used by other programs and/or that have existing instrumentation will also be used, such as the DOE Atmospheric Radiation Measurements (ARM) sites and the NASA Stennis satellite V&V site.

6. SUMMARY

MTI has a strong experimental program that combines the expertise of instrument scientists, remote sensing researchers, climate researchers, and facility modelers, as well as software and hardware developers. This experimental program would not be possible without the strong support of the US Department of Energy, not only financially, but also in developing teams and facilities at Los Alamos National Laboratory, Savannah River Technology Center, and Sandia National Laboratories and in encouraging the strong collaboration between these institutions. We expect other collaborations to emerge as the MTI launch approaches and research programs that use the imagery are developed.

REFERENCES

1. P.G. Weber, B.C. Brock, A.J. Garrett, B.W. Smith, C.C. Borel, W.B. Clodius, S.C. Bender, R.R. Kay and M.L. Decker, "Multispectral Thermal Imager mission overview," in this volume (1999).
2. C.C. Borel, W.B. Clodius, A. Davis, B.W. Smith, J.J. Szymanski, J. Theiler and P. Villeneuve, "MTI Core Science Retrieval Algorithms," in this volume (1999).
3. J. Theiler, S.P. Brumby, J.J. Szymanski, S. Alferink and J.J. Bloch, "Evolving Retrieval Algorithms with a Genetic Programming Scheme," in this volume (1999).
4. B.R. Stallard and J.G. Taylor, "Materials Identification and Anomaly Detection with Simulated MTI Data," in this volume (1999).
5. A.J. Garrett, R.J. Kurzeja, B.L. O'steen, E. Villa, M.J. Parker and M.M. Pendergast, "Post-launch Validation of Multispectral Thermal Imager (MTI) Data and Algorithms," in this volume (1999).
6. W.B. Clodius, S.C. Bender and B.W. Smith, "MTI on-orbit calibration," in this volume (1999).
7. R.G. Barry and A.L. Varani, Eds., *Earth Observing Science: Highlights 1994*, NSIDC Distributed Active Archive Center, Boulder, CO (1994).
8. R. Alexander and P. Cheatham, "Automated Hyperspectral Ground Processing," Proc. SPIE 3438 (1998).
9. R.R. Kay, T.D. Henson, J.L. Rienstra, M.L. Decker, N.G. Rackley, P.J. Claassen, R.E. Kidner, R.B. Taplin, D.M. Bullington, K.D. Marbach, C.E. Lanes, B.C. Brock, P.G. Weber, S.C. Bender, B.W. Smith, W.B. Clodius and C.C. Borel, "Multispectral Thermal Imager (MTI) Payload Overview," in this volume (1999).
10. HDF, 1999, Hierarchical Data Format standard, http://hdf.ncsa.uiuc.edu/, accessed May (1999).
11. J. Theiler, B.G. Henderson and B. W. Smith, "Algorithms using inter-band cross-correlation for pixel registration and jitter reconstruction in multi-channel push-broom imagers," Proc. SPIE 3163, 22 – 32 (1997).
12. J.J. Szymanski, C.C. Borel, Q.O. Harberger, P. Smolarkiewicz and J. Theiler, "Subpixel temperature retrieval with multispectral sensors," to appear in Proc. SPIE 3717 (1999).